Uludağ Üniversitesi

Bilgisayar Mühendisliği Tezli Yüksek Lisans Programı

Doğal Dil İşleme Dersi

1. Dönem – Proje 2

Hazırlayan

Nisanur BULUT – 501731014

Hedef

İngilizce tweet metinlerinin incelenerek, metinlerin kim tarafından yazıldığının bulunması hedeflenmiştir. Bu işlem için kullanılan ver seti, canlı olarak Twitter[1] sosyal medya platformundan sağlanmıştır.

Her tweet, tweet id, text, userid, location gibi pek çok nitelik tanımlayıcı nitelik içerir. Ancak analiz işlemi için yalnızca, tweet id ve tweet text'i kullanılmıştır. Proje içerisinde, analiz işlemi uygulandıkça çıkarım yapılan bilgiler bir sınıf nesnesi içinde tutulmuştur.

Projenin kodlanma aşamasında, python[2] programlama dili ve Jupyter Notebook [3] geliştirici ortamı kullanılmıştır.

Araştırma Öncesi Yapılan Çalışmalar

Tweet metinleri, kullanıcı temelli yapısız metinlerdir. Dolayısıyla duygusal ifadeler, emojilerin kullanımı, yazım ve noktalama işaretlerine uyulmaması ve dilbilgisi kurallarına uygun metin içeriklerinin Tweet içeriğinde olma ihtimali oldukça düşüktür.İçerikler yapısal değildir. Bu sebeple analiz işlemine başlamadan önce pek çok normalizasyon işlemi yapılmıştır. Bu işlem için, textacy[4] python kütüphanesinden yararlanılmıştır.

Textacy Kütüphanesi: SpaCy[5] kütüphanesi üzerine kurulu çeşitli doğal dil işleme görevlerini yerine getiren bir Python kütüphanesidir. Temel ilkeleri, tokenization, POS(part of speech tagger), bağımlılık, ayrıştırma vb. gibi işlemlerdir.

Yapılan Ön İşlemler

Ön işleme herhangi bir makine öğrenmesi veya derin öğrenme görevinden önce metin içerikli dökümanların hazırlanmasında en kritik adımdır.Tokenlara ayırma, gereksiz sık kullanılan kelimelerin(stop words) atılması ve kelime köklerinin bulunması en yaygın kullanılan ön işleme yöntemleridir.

Ana Adımlar :

- Bir metin dokümanını analiz etmek için, ilk olarak ***tokenlara ayırma işlemi*** yapılmalı ve kelime grupları elde edilmelidir.
- Tüm ortak ayırıcılar, işleçler, noktalama işaretleri ve yazdırılamayan karakterler ***kaldırılır***.
- Daha sonra, en sık kullanılan kelimeleri filtrelemeyi amaçlayan ***stop-words filtreleme*** gerçekleştirilir. Örnekler: "ama, belki, acaba".
- Son olarak, kelime hakkında dil-bilgisel veya sözcüksel bilgiler sunan son-eklerin çıkarılmasıyla morfolojik kökün elde edilmesini amaçlayan ***stemming*** ve / veya ***lemmatization*** uygulanır. Makalede, bu adımlar atlanmıştır.

Yapılan İşlemler

- Küçük harfe çevirme işlemi
- Url silme işlemi
- Email hesap silme işlemi
- Telefon numarası silme işlemi
- Para birimi silme işlemi

- Tırnak kullanımının kaldırılması işlemi
- Aksan içeren Unicode kullanımların kaldırılması işlemi

```
[ u'Excellent preliminary meeting in Oval with @SenSchumer - working on solutions for Security and our great Military together
with @SenateMajLdr McConnell and @SpeakerRyan. Making progress - four week extension would be best!'
 u'Just signed 702 Bill to reauthorize foreign intelligence collection. This is NOT the same FISA law that was so wrongly abus
ed during the election. I will always do the right thing for our country and put the safety of the American people first!'
 u'Today, I was honored and proud to address the 45th Annual @March_for_Life! You are living witnesses of this year\u2019s Mar
ch for Life theme: #LoveSavesLives. https://t.co/DMST4qhDmp'
 u'\u201cShutting down the government is a very serious thing. People die, accidents happen. I don\u2019t know how I would vot
e right now on a CR, OK?\u201d\nSen. Dianne Feinstein (D-Calif)\nhttps://t.co/7xP3CBnv5j'
 u'.@WhiteHouse Briefing with Director Marc Short and Director Mick Mulvaney...\nhttps://t.co/0O0VsYXmHB'
 u'Government Funding Bill past last night in the House of Representatives. Now Democrats are needed if it is to pass in the S
enate - but they want illegal immigration and weak borders. Shutdown coming? We need more Republican victories in 2018!'
 u'House of Representatives needs to pass Government Funding Bill tonight. So important for our country - our Military needs i
t!']
```

Resim1-Ön İşlem Öncesi

```
[u'excellent preliminary meeting in oval with senschumer working on solutions for security and our great military together wit
h senatemajldr mcconnell and speakerryan making progress four week extension would be best', u'just signed 702 bill to reautho
rize foreign intelligence collection this is not the same fisa law that was so wrongly abused during the election i will alway
s do the right thing for our country and put the safety of the american people first', u'today i was honored and proud to addr
ess the 45th annual marchforlife you are living witnesses of this years march for life theme lovesaveslives url', u'shutting d
own the government is a very serious thing people die accidents happen i dont know how i would vote right now on a cr ok\nsen
dianne feinstein dcalif\nurl', u'whitehouse briefing with director marc short and director mick mulvaney\nurl', u'government f
unding bill past last night in the house of representatives now democrats are needed if it is to pass in the senate but they w
ant illegal immigration and weak borders shutdown coming we need more republican victories in 2018', u'house of representative
s needs to pass government funding bill tonight so important for our country our military needs it']
```

Resim2-Ön işlem sonrası

Veri Gösterimi-Pandas Kütüphanesi

Twitter API kullanılarak elde edilen veriler yine bir Python kütüphanesi olan Pandas[6] kullanılarak kullanıcıya sunulmuştur. Pandas, veri analizi ve veri ön işlemeyi kolaylaştıran açık kaynak kodlu bir kütüphanedir. Dağıtık çalışmaya uygun değildir bu sebeple üzerinde işlem yapılan verinin büyüklüğü makinenin kapasitesiyle sınırlıdır özellikle de ana belleğin.

```
In [92]:
1 # A:
2 pd.DataFrame(hillary)
```

Out[92]:

	created_at	handle	mined_at	retweet_count	text	tweet_id
0	Fri Jan 19 19:08:40 +0000 2018	HillaryClinton	2018-01-20 02:08:40.529365	16059	I'm so heartened by all of you. Onward! https:...	954430425321046016
1	Mon Jan 15 20:48:09 +0000 2018	HillaryClinton	2018-01-20 02:08:40.529391	23021	These words from Dr. King also come to mind to...	953005910930153472
2	Mon Jan 15 17:38:41 +0000 2018	HillaryClinton	2018-01-20 02:08:40.529402	8990	Beautifully said, @BerniceKing. An important m...	952958227825782784
3	Fri Jan 12 19:14:45 +0000 2018	HillaryClinton	2018-01-20 02:08:40.529413	59635	The anniversary of the devastating earthquake ...	951895239140298752
4	Fri Jan 12 03:07:54 +0000 2018	HillaryClinton	2018-01-20 02:08:40.529421	320	@NancyEMcFadden Nancy has a record of beating ...	951651923328987136
5	Tue Jan 02 17:27:52 +0000 2018	HillaryClinton	2018-01-20 02:08:40.529430	8529	Families across America had to start 2018 worr...	948244463138328577
6	Tue Jan 02 17:26:39 +0000 2018	HillaryClinton	2018-01-20 02:08:40.529438	14714	Time to bring CHIP to the Senate floor as prom...	948244159986651136
7	Sun Dec 31 03:49:33 +0000 2017	HillaryClinton	2018-01-20 02:08:40.529447	20666	The Iranian people, especially the young, are ...	947313751992274944

Resim 3-Pandas Kütüphanesi Tablo Kullanımı

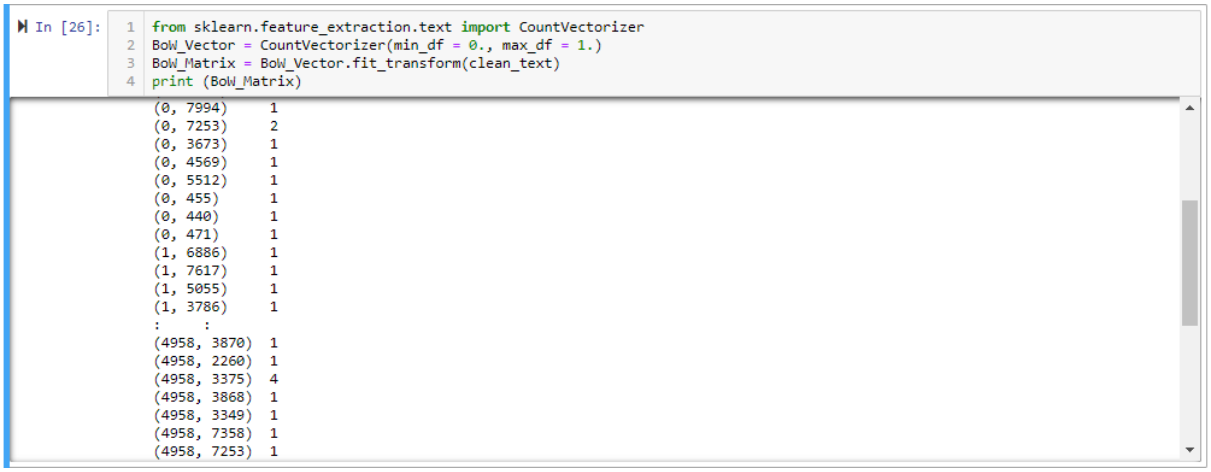
Analiz İşlemi-Vektör Kullanımı

Projenin gerçekleştirilmesi aşamasında kullanılan veri seti Tweet metinleridir. Tweet metinleri yapısal metinler değildir. Yapısal olmayan bu metinler üzerinde makine öğrenmesi modellerinin uygulanabilmesi için öncelikle metinlerin işlenmesi gerekmektedir. Kabaca şu adımlar izlenir, metin içerikleri özniteliklere yani yapısal bir formata çevrilir, öznitelikler ise vektörlere çevrilir.

Öznitelik/Terim Temsili ve BoW Modeli

Kelime/Sözcük Çantası (BoW) modeli, bir dokümandaki terimlerin oluşum şeklini (Örneğin: terim sayılarını) belirten metnin temsil biçimidir. Bu modelde; terim pozisyonu ve sözcük sıralaması dikkate alınmaz.

Vektör Uzay Modeli (VUM), her bir metin dokümanının bir vektör olarak temsil edildiği gelişmiş bir BoW sürümüdür ve her bir boyut ayrı bir terime (özniteliğe) karşılık gelir. Dokümanda bir terim yer almışsa, ilgili doküman vektöründe terimin değeri sıfırdan farklı olur.



Resim 4- 2.Kullanıcının Tweet Metninin BoW Model Gösterimi

TF x IDF Skorlama Modeli

Terim sıklıklarının sayılması ile ilgili en önemli sorun, sık kullanılan terimlerin dokümanda baskın olmaları ve artık dokümanı temsil eder hale gelmeleridir. Bu terimler çok değerli bilgiler içermese dahi özellik kümesindeki diğer niteliklerin etkisiz olmasına sebep olur.

Bu problemi çözmek için, "Terim Frekansı x Ters Belge Frekansı" anlamına gelen "TF x IDF" modelini ve skorlama yöntemi kullanılabilir. Hesaplamada iki ölçüt kullanılır*:* terim sıklığı (tf) *ve* ters belge sıklığı (idf). *TF x* IDF'nin matematiksel denklemleri şöyledir:

- j'ninci dokümandaki i'ninci terim için TF x IDF skoru = TF(i, j) * IDF(i)
- TF(i, j) = (Dokümandaki i'ninci terimin sıklığı) / (Dokümandaki toplam terim sayısı)
- IDF(i) = log2(Toplam doküman sayısı/i'ninci terimi içeren doküman sayısı)

❖ Kodlama adımında dokümanlarımızı bir TFxIDF özellik matrisine dönüştürebilmek için TfidfVectorizer sınıfı kullanılmıştır.

Konu Modelleme

Tweet metninin kime ait olduğunu anlayabilmek için, metin içeriği üzerinde modelleme işlemi yapılır. Bu işlem için metin içeriğindeki anahtar kelimelerin bulunup ortaya çıkarılması gerekmektedir. Çıkartılan konular bir terim koleksiyonu olarak temsil edilirler. Metin dokümanlarının büyük bir kısmını özetlemek için çok değerlidirler, dahası, verilerde gizli kalıpları/anlamsallığı ortaya çıkarırlar.

Metin içeriklerinin birbirine olan benzerliğini kontrol etmek için LDA (Latent Dirichlet Allocation) [7] modeli kullanılır.

LDA: Verilerdeki bazı bölümlerin neden benzer olduğunu tarif etmek amacıyla, gözlem gruplarının gözlemlenmemiş gruplar tarafından açıklanmasına izin veren bir üretken istatistik modelidir.

Python'da gensim[8] ve sklearn kütüphanelerini kullanarak LDA modelinin oluşturulması mümkündür. Proje kapsamında sklearn kütüphanesi kullanılmıştır.

NGram Algoritması

Mevcut veri seti kontrol edilerek, yazarı belli olmayan tweet metinlerine dair yazar tahmini yapmak hedefine ulaşmak için, kontrol edilmek istenen verinin eldeki veriye olan yakınlığının hesaplanması sürecinde sklearn kütüphanesinin ngram sınıfından yararlanılır.

Ngram aslında bir dil modelidir. Tahmine ve olasılığa dayanır. Karakter ve kelime bazlı olmak üzere iki kategoride incelenir. Bir karakter ya da kelimenin kendinden önce gelen birkaç karaktere ya da kelimeye bağlı olarak, bulunduğu yerde olma olasılığını hesaplar.

- Projede ngram karakter temelinde kullanılmıştır.

Logistic Regression

Yalnızca iki değere sahip veriseti üzerinde olaılık hesaplama işlemidir. Veri setinde her kelimenin ağırlıkları hesaplanmıştır. Bu ağırlıklı veri üzerinde lojistik resgresyon işlemi yapılır.

Model Seçimi-GridSearchCV

Eğitim veri seti üzerinde yapılan önişleme, ngram ağaç oluşumu, matris ve vektör dönüşümlerinin ile lojistik regresyon işleminin ardından model seçimi aşamasına geçilir. Model seçiminde, eşlemenin kullanılacağı algoritmanın başarısını yükseltmek için en iyi parametrelerin belirlenmesi gibi işlemler için Python Sklearn kütüphanesinin GridSearchCV sınıfından yararlanılmıştır.

Çapraz eşleme işlemleriyle birlikte eğitim seti 10 ile çarpılıp bölünür. Böylece 200 tweet metni içinde elde 2000 veri bulunur. Bu veri üzerinde çalışacak algoritma için gerekli en iyi parametreler elde edilmiş olunur.

```
In [18]:
from sklearn.model_selection import GridSearchCV

#parametre ayarı yaparak başarının yükselmesi amacıyla GridSearchCV kullanılmıştır

lr = LogisticRegression()
params = {'penalty': ['l1', 'l2'], 'C':np.logspace(-5,0,100)}
#Grid searching to find optimal parameters for Logistic Regression
gs = GridSearchCV(lr, param_grid=params, cv=10, verbose=1)
gs.fit(X, y)

Fitting 10 folds for each of 200 candidates, totalling 2000 fits

[Parallel(n_jobs=1)]: Done 2000 out of 2000 | elapsed:  4.5min finished

Out[18]: GridSearchCV(cv=10, error_score='raise',
       estimator=LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
          penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
          verbose=0, warm_start=False),
       fit_params=None, iid=True, n_jobs=1,
       param_grid={'penalty': ['l1', 'l2'], 'C': array([1.00000e-05, 1.12332e-05, ..., 8.90215e-01, 1.00000e+00])},
       pre_dispatch='2*n_jobs', refit=True, return_train_score='warn',
       scoring=None, verbose=1)
```

Resim 5-Model Seçimi

Uygulama

Projenin uygulanması aşamasında, iki adet tweet metni uygulamaya girdi olarak verilir. Her metnin 2 kullanıcıdan hangisine ait olduğu olasılık değerleriyle gösterilir.

- Metinler test işlemi için öncelikle Tfidf vektör yapısına getirilir.

In [21]:
```
estimator = LogisticRegression(penalty='l2',C=1.0)
estimator.fit(X,y)

# Kaynağın TfIdf vector olarak ayarlanması
source_test = [
    "The presidency doesn't change who you are—it reveals who you are. And we've seen all we need to of Donald Trump.",
    "Crooked Hillary is spending tremendous amounts of Wall Street money on false ads against me. She is a very dishonest pers
]

Xtest = tfv.transform(source_test)
pd.DataFrame(estimator.predict_proba(Xtest), columns=["Proba_Hillary", "Proba_Trump"])
```

Out[21]:

	Proba_Hillary	Proba_Trump
0	0.906631	0.093369
1	0.277569	0.722431

Resim 6-Uygulama 1

In [36]:
```
# Kaynağın TfIdf vector olarak ayarlanması
source_test = [
    "To all the little girls watching...never doubt that you are valuable and powerful & deserving of every chance & opportunit
    "Nevada just became the first state in the country to have a legislature that's majority women. Let's make it the first of
]

Xtest = tfv.transform(source_test)
pd.DataFrame(estimator.predict_proba(Xtest), columns=["Proba_Hillary", "Proba_Trump"])
```

Out[36]:

	Proba_Hillary	Proba_Trump
0	0.742282	0.257718
1	0.850053	0.149947

In [35]:
```
# Kaynağın TfIdf vector olarak ayarlanması
source_test = [
    "I am all alone (poor me) in the White House waiting for the Democrats to come back and make a deal on desperately needed B
    "Saudi Arabia has now agreed to spend the necessary money needed to help rebuild Syria, instead of the United States. See?
]

Xtest = tfv.transform(source_test)
pd.DataFrame(estimator.predict_proba(Xtest), columns=["Proba_Hillary", "Proba_Trump"])

```

Out[35]:

	Proba_Hillary	Proba_Trump
0	0.074154	0.925846
1	0.288556	0.711444

Resim 7-Uygulama 2

:
```
# Kaynağın TfIdf vector olarak ayarlanması
source_test = [
    "The Wall is different than the 25 Billion Dollars in Border Security. The complete Wall will be built with the Shutdown mo
    "There's new CDC data out about gun deaths in America. Last year saw the most gun deaths in 40 years, nearly 40,000 people.
]

Xtest = tfv.transform(source_test)
pd.DataFrame(estimator.predict_proba(Xtest), columns=["Proba_Hillary", "Proba_Trump"])
```

:

	Proba_Hillary	Proba_Trump
0	0.055058	0.944942
1	0.726158	0.273842

Resim 8-Uygulama 3

Eldeki tüm tweet metinleri birleştirilir ve Model kullanılarak hangi kullanıcıya ait olduğu test edilir. Tweet metninin gerçekten kime ait olduğu handle sütununda gösterilmiştir. Model kullanımının ardından kime ait olduğunun olasılıklı değerleri ise tablonun son iki sütununda gösterilmiştir. Bu test işleminin amacı modelin ne derece doğru çalıştığını gözlemlemektir.

```
In [23]: 1 Probas_x = pd.DataFrame(estimator.predict_proba(X), columns=["Proba_Hillary", "Proba_Donald"])

In [24]: 1 joined_x = pd.merge(tweets, Probas_x, left_index=True, right_index=True)

In [25]: 1 joined_x
```

Out[25]:

	created_at	handle	mined_at	retweet_count	text	tweet_id	Proba_Hillary	Proba_Donald
0	Mon Dec 24 15:55:22 +0000 2018	realDonaldTrump	2018-12-24 19:43:19.714021	7771	The only problem our economy has is the Fed. T...	1077231267559755776	0.293076	0.706924
0	Fri Dec 21 18:05:19 +0000 2018	HillaryClinton	2018-12-24 19:43:57.412145	745	As we finalize our plans for 2019 and beyond, ...	1076176803826421761	0.293076	0.706924
1	Mon Dec 24 15:33:41 +0000 2018	realDonaldTrump	2018-12-24 19:43:19.714520	8738	AMERICA IS RESPECTED AGAIN!	1077225810329825281	0.276149	0.723851
1	Fri Dec 21 18:05:18 +0000 2018	HillaryClinton	2018-12-24 19:43:57.412145	679	There was a 10-point increase in youth turnout...	1076176803126001665	0.276149	0.723851
2	Mon Dec 24 15:23:22 +0000 2018	realDonaldTrump	2018-12-24 19:43:19.714520	9682	For all of the sympathizers out there of Brett...	1077223213598429185	0.431758	0.568242
2	Fri Dec 21 18:05:18 +0000 2018	HillaryClinton	2018-12-24 19:43:57.412645	275	Five @EmergeAmerica alumni are headed to Congr...	1076176802387759104	0.431758	0.568242
3	Mon Dec 24 14:59:23 +0000 2018	realDonaldTrump	2018-12-24 19:43:19.714520	10225	...We are substantially subsidizing the Milit...	1077217178322194432	0.272501	0.727499
3	Fri Dec 21 18:05:18 +0000 2018	HillaryClinton	2018-12-24 19:43:57.412645	306	For many campaigns, @SwingLeft was the wave. O...	1076176801674772483	0.272501	0.727499
4	Mon Dec 24 14:41:02 +0000 2018	realDonaldTrump	2018-12-24 19:43:19.715020	10444	To those few Senators who think I don't like o...	1077212559604924416	0.320071	0.679929
4	Fri Dec 21 18:05:18 +0000 2018	HillaryClinton	2018-12-24 19:43:57.412645	251	In 2018, @LatinoVictoryUS increased their endo...	1076176800676569088	0.320071	0.679929
5	Mon Dec 24 14:31:50 +0000 2018	realDonaldTrump	2018-12-24 19:43:19.715020	11922	Virtually every Democrat we are dealing with t...	1077210242574942208	0.301271	0.698729

Resim9- Uygulama 4

Birleştirilmiş olan tweet metinlerinde, her iki kullanıcı için olasılık değeri en yüksek olan tweet metinleri resim 10'da gösterilmiştir.

```
In [26]:
#hillary clinton'a ait olabilecek en güçlü aday tweet metni
joined_hillary = joined_x[joined_x['handle']=="HillaryClinton"]
for el in joined_hillary[joined_hillary['Proba_Hillary']==max(joined_hillary['Proba_Hillary'])]['text']:
    print (el)
```

```
"The place of your birth should never be a barrier." A letter to young undocumented people on DACA's anniversary: https://t.co/4dDrLvefeM
```

```
In [27]:
#hillary clinton'a ait olabilecek en düşük olasılıklı aday tweet metni
for el in joined_hillary[joined_hillary['Proba_Hillary']==min(joined_hillary['Proba_Hillary'])]['text']:
    print (el)
```

```
"He thinks we should be afraid of our Muslim brothers and sisters—because he has no idea who they really are." —@FLOTUS on Trump
```

```
In [28]:
#DonaldTrump'a ait olabilecek en güçlü aday tweet metni
joined_donald = joined_x[joined_x['handle']=="realDonaldTrump"]
for el in joined_donald[joined_donald['Proba_Donald']==max(joined_donald['Proba_Donald'])]['text']:
    print (el)
```

```
THE FAKE NEWS MEDIA IS THE OPPOSITION PARTY. It is very bad for our Great Country....BUT WE ARE WINNING!
```

```
In [29]:
#DonaldTrump'a ait olabilecek en düşük olasılıklı aday tweet metni
for el in joined_donald[joined_donald['Proba_Donald']==min(joined_donald['Proba_Donald'])]['text']:
    print (el)
```

```
Statement on National Strategy for Counterterrorism: https://t.co/ajFBg9Elsj https://t.co/Qr56ycjMAV
"USMCA Wins Praise as a Victory for American Industries and Workers"➡https://t.co/UWhPhNXiiL https://t.co/eBVv0nXXRT
#HurricaneFlorence
https://t.co/mP7icn0Yzl https://t.co/jOdKT02rbH
#500Days of American Greatness: https://t.co/qQEEPqmaax https://t.co/1VCeoTk1cI
```

Resim 10- Uygulama 5

Kaynakça

1. https://shiftdelete.net/twitter-nedir-nasil-kullanilir-10080
2. https://www.python.tc/python-nedir/
3. http://www.veridefteri.com/2017/10/30/jupyter-notebook-nedir-2/
4. https://pypi.org/project/textacy/
5. https://pypi.org/project/spacy/
6. http://www.datascience.istanbul/2017/05/21/python-pandas-ile-temel-islemler-1/
7. https://scikitlearn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html
8. https://radimrehurek.com/gensim/